Modeling Enrollment Disparities Across the Postsecondary Admissions Pipeline

Amrit Singh

January 22, 2026

Statistical Models for Data Science (RS3)

Thomas Jefferson High School for Science and Technology

Lit Review/ Introduction:

Access to postsecondary education is a key determinant of long-term socioeconomic outcomes, including lifetime earnings, health, and intergenerational mobility. Despite the expansion of higher education over recent decades, substantial disparities in college access and enrollment persist across demographic and geographic contexts. Understanding where students exit the postsecondary pipeline, whether at the application, admission, or enrollment stage, is critical for identifying structural barriers that limit equitable access to higher education.

The college admissions process is best understood as a multi-stage pipeline. Students first apply to institutions, a subset of applicants are admitted, and an even smaller subset ultimately enrolls. Attrition at any stage can substantially affect final enrollment outcomes. While prior research has examined these stages individually, fewer studies model how earlier pipeline stages jointly predict enrollment outcomes within a single statistical framework. As a result, it remains unclear whether enrollment disparities are driven primarily by differences in applicant volume, admissions decisions, or post-admission enrollment behavior.

A substantial body of literature has documented widening socioeconomic disparities in educational attainment. Reardon (2011) demonstrates that income-based achievement gaps have grown significantly over time, surpassing racial achievement gaps in magnitude. These disparities shape not only academic preparation but also access to information, financial resources, and institutional support necessary for navigating the college application process. Students from lower-income backgrounds are therefore less likely to apply to college and, even when admitted, may face barriers to enrollment.

Institutional practices further influence admissions outcomes. Posselt et al. (2012) show that admissions criteria and institutional priorities at selective universities can disproportionately disadvantage certain groups, even among academically qualified applicants. Their findings highlight that admissions outcomes reflect not only student characteristics but also institutional decision-making processes. However, this work focuses primarily on admissions decisions and does not fully go into whether admitted students ultimately enroll.

Post-admission factors also play a critical role in shaping enrollment outcomes. Hoxby and Avery (2013) document the phenomenon of "undermatching," where high-achieving students, particularly those from disadvantaged backgrounds, either do not apply to selective institutions or fail to enroll after admission. Their work suggests that enrollment disparities persist even when admissions probabilities are comparable, emphasizing the importance of examining post-admission behavior.

Collectively, prior research demonstrates that disparities in postsecondary access emerge across multiple stages of the admissions pipeline. However, most studies analyze application behavior, admissions decisions, or enrollment outcomes independently. Few studies quantitatively model how applicant counts, admission counts, and enrollment outcomes interact simultaneously using a unified dataset. This study addresses this by applying multiple linear regression to model enrollment outcomes as a function of earlier pipeline stages, providing a clearer understanding of where disparities in postsecondary access are most concentrated.

Research Question: To what extent do differences in application volume, admission outcomes, and post-admission yield predict enrollment totals across counties and over time within the postsecondary education pipeline?

Data & Sample Description

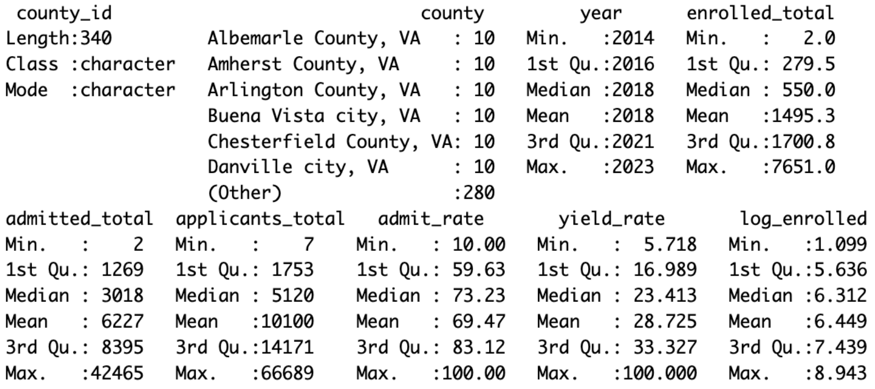
The dataset used in this analysis consists of county-year observations capturing key stages of the postsecondary admissions pipeline. Each observation represents a single county in a given year and includes information on applicant counts, admissions totals, enrollment outcomes, and associated rates. This structure allows the analysis to examine both geographic variation across counties and changes over time.

After cleaning, the final analytic sample includes 340 county-year observations spanning the years 2014 through 2023. Observations with missing values in any primary admissions pipeline variables were excluded to ensure consistency across regression models.

The variable county represents a geographic aggregation corresponding to individual counties or independent cities. Counties are included as fixed effects in the regression models to control for unobserved, time-invariant geographic characteristics (such as population size, institutional density, or regional socioeconomic conditions) that may influence enrollment outcomes but are not directly measured in the dataset. Including county fixed effects allows the analysis to isolate within-county associations between admissions pipeline variables and enrollment outcomes over time.

Summary statistics indicate substantial variation across counties and years. Enrollment totals range from very small values in lower-population counties to several thousand students in higher-population counties. Applicant and admission counts are right-skewed, with mean values exceeding medians, while admission and yield rates also exhibit meaningful variability. This heterogeneity motivates a multivariable modeling approach that jointly considers multiple stages of the admissions pipeline.

Summary Statistics of Key Admissions Pipeline Variables



```
  county_id                              county        year       enrolled_total
 Length:340         Albemarle County, VA    : 10   Min.   :2014   Min.   :   2.0
 Class :character   Amherst County, VA      : 10   1st Qu.:2016   1st Qu.: 279.5
 Mode  :character   Arlington County, VA    : 10   Median :2018   Median : 550.0
                    Buena Vista city, VA    : 10   Mean   :2018   Mean   :1495.3
                    Chesterfield County, VA: 10   3rd Qu.:2021   3rd Qu.:1700.8
                    Danville city, VA       : 10   Max.   :2023   Max.   :7651.0
                    (Other)                 :280
 admitted_total  applicants_total   admit_rate       yield_rate       log_enrolled
 Min.   :    2   Min.   :    7    Min.   : 10.00   Min.   :  5.718   Min.   :1.099
 1st Qu.: 1269   1st Qu.: 1753    1st Qu.: 59.63   1st Qu.: 16.989   1st Qu.:5.636
 Median : 3018   Median : 5120    Median : 73.23   Median : 23.413   Median :6.312
 Mean   : 6227   Mean   :10100    Mean   : 69.47   Mean   : 28.725   Mean   :6.449
 3rd Qu.: 8395   3rd Qu.:14171    3rd Qu.: 83.12   3rd Qu.: 33.327   3rd Qu.:7.439
 Max.   :42465   Max.   :66689    Max.   :100.00   Max.   :100.000   Max.   :8.943
```

Data Cleaning & Variable Construction

To prepare the dataset for analysis, several preprocessing steps were performed to improve interpretability and ensure consistency across observations. Variable names were first standardized and renamed to clearly reflect their roles in the admissions pipeline. For example, original column names were replaced with more interpretable labels such as applicants_total, admitted_total, enrolled_total, admit_rate, and yield_rate, facilitating clearer interpretation of regression results.

Rows containing missing values in any of the key admissions pipeline variables were removed prior to analysis. Because the regression models require complete information across all predictors and the response variable, excluding incomplete observations ensured that estimates were based on a consistent set of county-year observations and avoided complications associated with imputation in a multivariable setting.

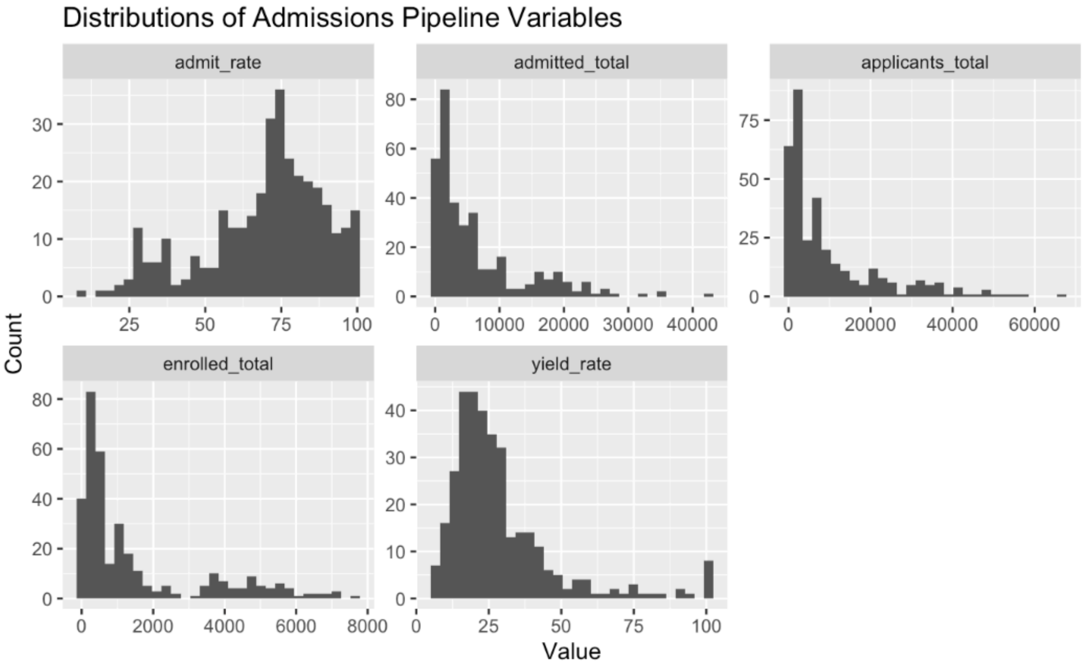
The variable county was treated as a categorical factor to allow for the inclusion of county fixed effects in the regression models. Modeling county as a factor accounts for unobserved, time-invariant differences across counties, such as long-standing socioeconomic conditions or institutional characteristics, that may influence enrollment outcomes but are not explicitly measured in the dataset.

The response variable for all models is enrolled_total, representing the total number of students who ultimately enrolled in postsecondary institutions within a county in a given year. Predictor variables include applicants_total, capturing application volume; admitted_total, reflecting admissions decisions; admit_rate and yield_rate, representing proportional outcomes at different stages of the pipeline; year, accounting for temporal trends; and county, included as a categorical fixed effect. Together, these variables allow for a comprehensive examination of how different stages of the admissions pipeline jointly relate to enrollment outcomes.

Exploratory Data Analysis (EDA)

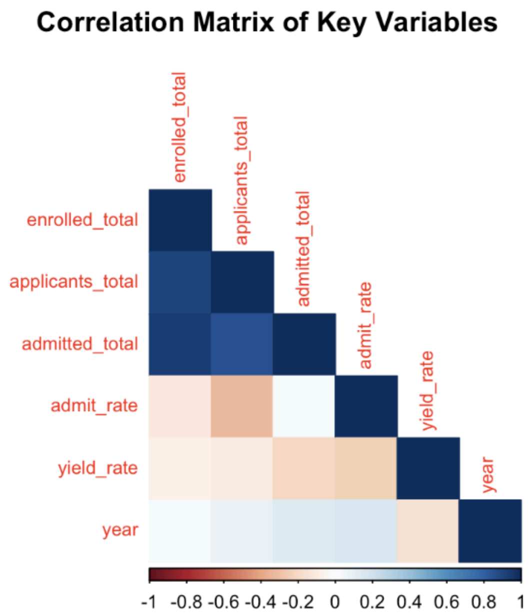
Exploratory data analysis was conducted to examine the distributions of key admissions pipeline variables and to assess relationships among variables prior to formal modeling. This step provides insight into the underlying structure of the data and informs later decisions regarding model specification and diagnostic checks.

The distributions of the primary admissions pipeline variables are shown in Figure 1. Count variables (applicants_total, admitted_total, and enrolled_total) exhibit pronounced right skew, with most county–year observations concentrated at relatively low values and a small number of counties accounting for very large volumes. This pattern reflects substantial heterogeneity in population size and postsecondary participation across counties. In contrast, the proportional variables admit_rate and yield_rate display broader but bounded distributions, indicating meaningful variation in both admissions selectivity and post-admission enrollment behavior across counties and years. The skewness observed in the count variables motivates evaluation of linearity and variance assumptions in subsequent regression diagnostics.



Relationships among the admissions pipeline variables are summarized in Figure 2, which presents a correlation matrix of key variables. Strong positive correlations are observed among the count variables (applicants_total, admitted_total, and enrolled_total), reflecting the sequential nature of the admissions process. Admission and yield rates exhibit weaker correlations with raw counts, suggesting that proportional outcomes capture distinct information beyond application volume alone. Year is only modestly correlated with other variables, indicating that temporal trends exist but do not dominate cross-sectional variation.

These correlation patterns support the use of a multivariable regression framework to jointly model multiple stages of the admissions pipeline. At the same time, the presence of strong correlations among predictors motivates explicit checks for multicollinearity in later modeling stages using variance inflation factors (VIF).



Model Specification

The primary objective of this analysis is to model variation in postsecondary enrollment outcomes as a function of earlier stages in the admissions pipeline. To do so, a multiple linear regression framework is used to estimate how application volume, admissions outcomes, and post-admission behavior jointly relate to enrollment totals across counties and over time.

Formally, enrollment totals are modeled as a function of key admissions pipeline variables, temporal trends, and county fixed effects. In words, the model predicts the total number of enrolled students in a given county and year based on the number of applicants, the number admitted, the admission rate, the yield rate, the year of observation, and a set of county-specific indicator variables. This specification allows the analysis to assess the relative contribution of different pipeline stages while controlling for persistent geographic differences.

Enrollment totals (enrolled_total) are chosen as the outcome of interest because they represent the final realized outcome of the admissions process. While application and admission counts capture intermediate stages of access, enrollment reflects whether students ultimately matriculate, making it the most policy-relevant measure of postsecondary participation. Modeling enrollment directly allows the analysis to identify which stages of the pipeline are most strongly associated with final enrollment outcomes.

Each admissions pipeline variable is included to capture a distinct component of the process. Applicants_total reflects demand for postsecondary education and access to the application stage. Admitted_total captures institutional admissions decisions, while admit_rate reflects selectivity conditional on applicant volume. Yield_rate represents post-admission behavior, capturing the proportion of admitted students who ultimately enroll. Including both count-based and rate-based measures allows the model to distinguish between volume effects and proportional differences across pipeline stages.

The variable year is included to account for temporal trends that may influence enrollment outcomes across all counties, such as changes in demographic composition, economic conditions, or broader shifts in postsecondary participation over time. Controlling for year ensures that estimated relationships between pipeline variables and enrollment are not confounded by common time-related shocks.

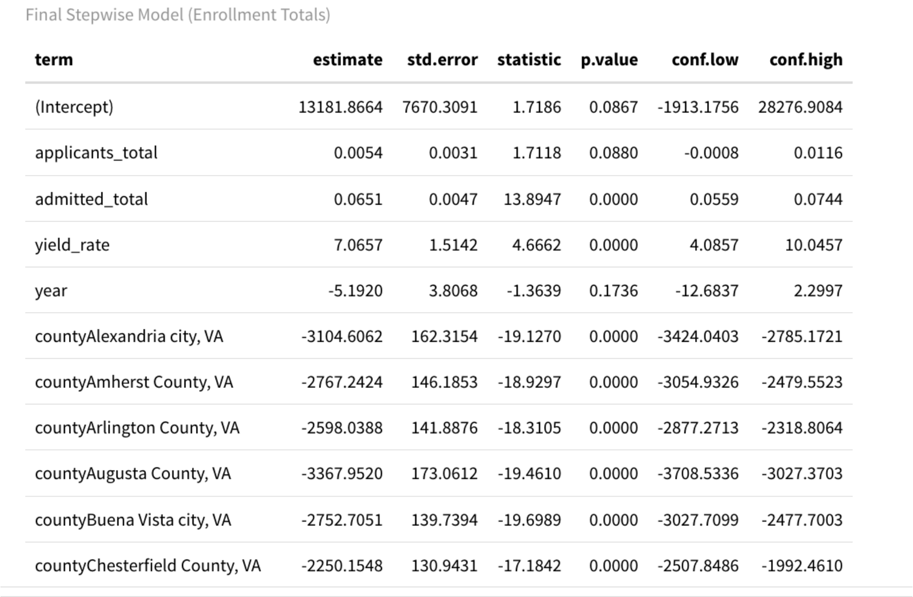
Finally, county is included as a set of fixed effects to control for unobserved, time-invariant differences across geographic areas. These fixed effects account for persistent county-level characteristics (such as population size, regional economic conditions, or long-standing educational infrastructure) that may influence enrollment outcomes but are not directly measured in the dataset. Including county fixed effects allows the model to focus on within-county variation over time, strengthening the interpretation of estimated associations between admissions pipeline variables and enrollment totals.

Model Selection & Multicollinearity

To identify a parsimonious yet explanatory model, I used AIC-based stepwise regression starting from the full specification that included all admissions pipeline variables, year, and county fixed effects. The Akaike Information Criterion (AIC) balances model fit against complexity, allowing the selection procedure to retain predictors that meaningfully improve explanatory power while penalizing unnecessary parameters.

The stepwise procedure converged on a final model that retains applicants_total, admitted_total, yield_rate, year, and county fixed effects. This result reflects both statistical significance and conceptual relevance: enrollment is mechanically linked to admitted students and yield rates, while county and year capture persistent geographic and temporal structure not explained by pipeline counts alone.

This table presents the coefficient estimates, standard errors, and confidence intervals for the final stepwise model predicting total enrollment.



Final Stepwise Model (Enrollment Totals)

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	13181.8664	7670.3091	1.7186	0.0867	-1913.1756	28276.9084
applicants_total	0.0054	0.0031	1.7118	0.0880	-0.0008	0.0116
admitted_total	0.0651	0.0047	13.8947	0.0000	0.0559	0.0744
yield_rate	7.0657	1.5142	4.6662	0.0000	4.0857	10.0457
year	-5.1920	3.8068	-1.3639	0.1736	-12.6837	2.2997
countyAlexandria city, VA	-3104.6062	162.3154	-19.1270	0.0000	-3424.0403	-2785.1721
countyAmherst County, VA	-2767.2424	146.1853	-18.9297	0.0000	-3054.9326	-2479.5523
countyArlington County, VA	-2598.0388	141.8876	-18.3105	0.0000	-2877.2713	-2318.8064
countyAugusta County, VA	-3367.9520	173.0612	-19.4610	0.0000	-3708.5336	-3027.3703
countyBuena Vista city, VA	-2752.7051	139.7394	-19.6989	0.0000	-3027.7099	-2477.7003
countyChesterfield County, VA	-2250.1548	130.9431	-17.1842	0.0000	-2507.8486	-1992.4610

The inclusion of multiple pipeline stages in the final model highlights the importance of modeling enrollment as the outcome of a sequential admissions process rather than a single-input system. While applicants_total shows weaker marginal significance once admitted_total is included, it remains informative in capturing variation in the size of the applicant pool across county-years. Yield_rate remains strongly predictive, reinforcing its role as the final conversion stage translating offers into enrolled students.

County fixed effects are consistently large and statistically significant, indicating substantial baseline differences in enrollment levels across counties that persist after controlling for admissions dynamics and time trends.

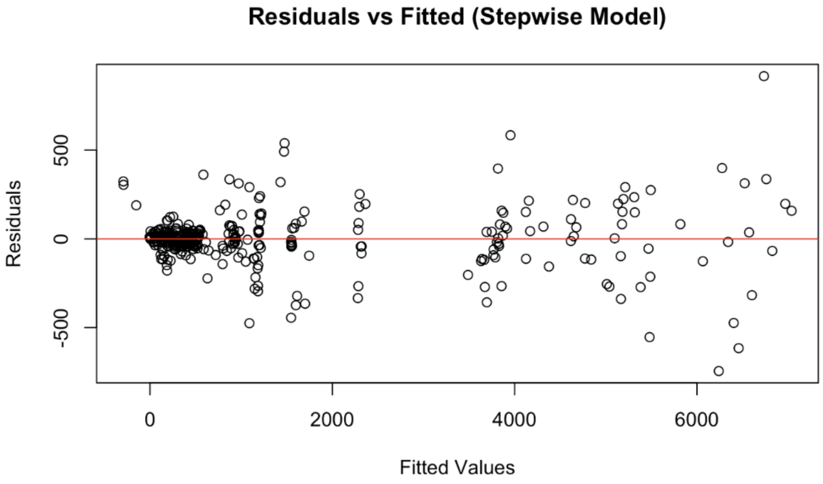
Given the strong correlations observed during exploratory analysis (particularly among applicants_total, admitted_total, and enrolled_total) multicollinearity was explicitly evaluated using Variance Inflation Factors (VIFs) computed from the final stepwise model.

VIF diagnostics indicated elevated but manageable multicollinearity among pipeline variables, which is expected due to their structural relationship within the admissions process. Importantly, no VIF values exceeded conventional thresholds that would indicate severe instability or invalidate coefficient interpretation. Rather than removing theoretically essential predictors, VIF diagnostics were used as a monitoring tool to ensure coefficient estimates remained numerically stable and interpretable.

Overall, the combination of AIC-based selection and VIF diagnostics supports the final model as a statistically sound and substantively meaningful representation of enrollment dynamics across counties and years.

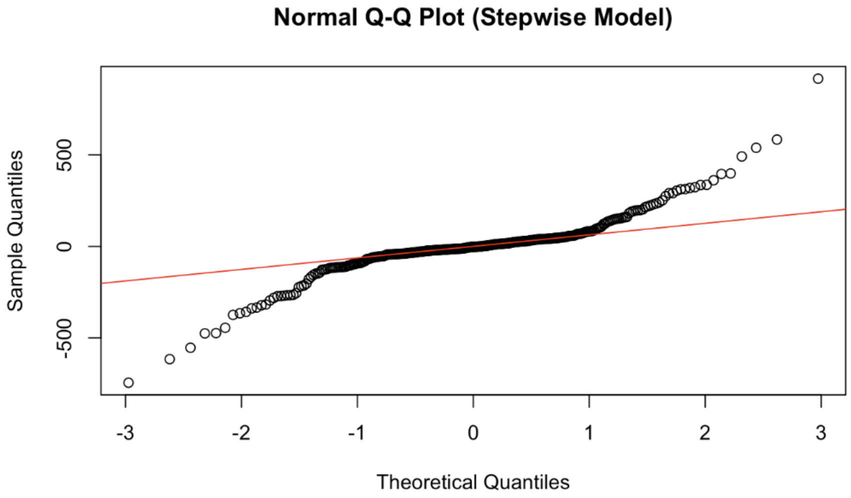
Model Diagnostics & Assumption Checks

This section evaluates whether the assumptions underlying the linear regression model are reasonably satisfied. Diagnostic plots and formal tests are used to assess linearity, constant variance, normality of errors, and independence.

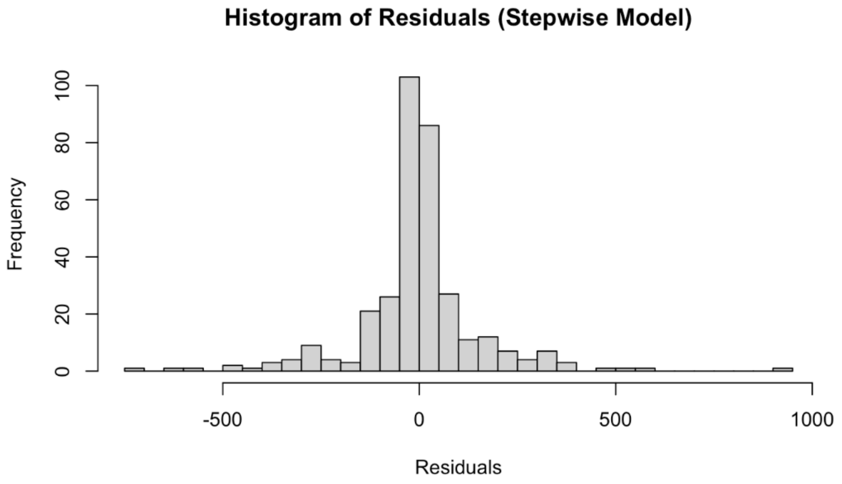


The residuals versus fitted values plot shows no strong systematic pattern, supporting the assumption of linearity in the relationship between predictors and enrollment totals. Residuals are centered around zero across the range of fitted values, indicating that the model does not consistently over- or under-predict enrollment at particular levels.

There is some increase in the spread of residuals at higher fitted values, suggesting mild heteroskedasticity. This pattern is not unexpected given the right-skewed nature of enrollment counts and the presence of counties with substantially larger enrollment totals. While variance is not perfectly constant, the deviation is moderate rather than severe.



The Q-Q plot indicates that residuals are approximately normally distributed, particularly in the central portion of the distribution. Minor deviations are observed in the tails, with some extreme positive and negative residuals. Given the relatively large sample size and the count-based nature of the outcome variable, these departures are considered acceptable and unlikely to meaningfully bias inference.

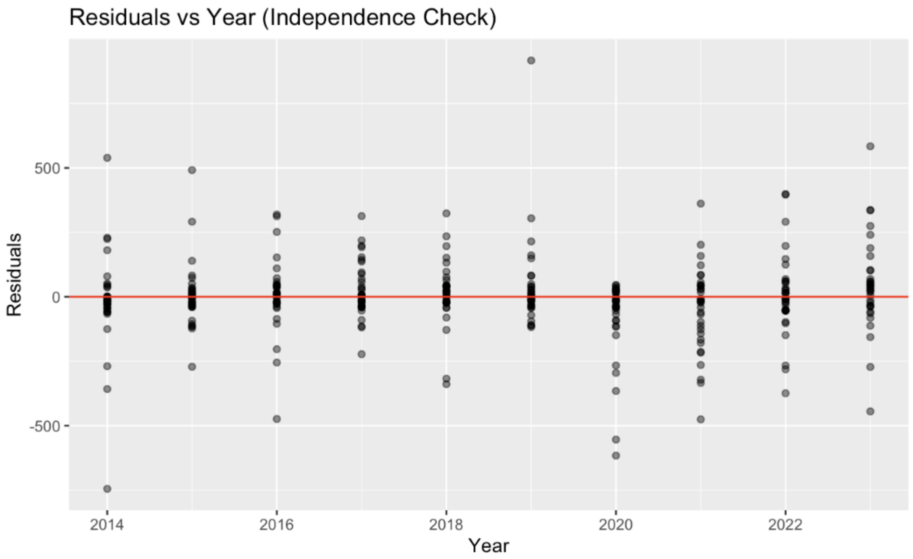


The histogram further supports approximate normality, showing a symmetric distribution centered near zero with slightly heavier tails. With the Q–Q plot, this suggests the normality assumption is reasonably satisfied for estimation and hypothesis testing.

Formal Tests for Heteroskedasticity

To evaluate constant variance, a Breusch–Pagan test was conducted on the stepwise model. The test provides evidence of heteroskedasticity, consistent with the increasing residual spread observed in the residuals versus fitted plot.

Because heteroskedasticity can lead to underestimated standard errors and inflated t-statistics, HC1 heteroskedasticity-robust standard errors were used when reporting coefficient inference. This adjustment ensures that statistical conclusions remain valid even when the constant variance assumption is violated.

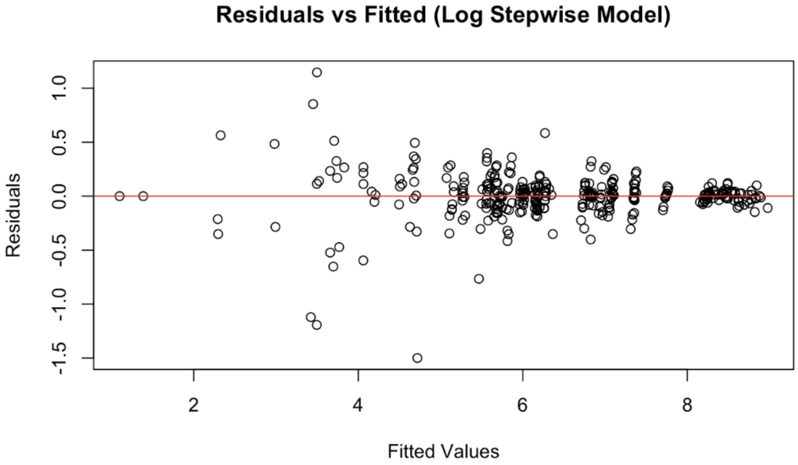


The residuals plotted against year show no strong temporal pattern or systematic trend. Residuals remain centered around zero across the time span of the data, with no evidence of serial correlation or year-specific clustering. This supports the assumption that errors are independent across observations after controlling for year and county fixed effects.

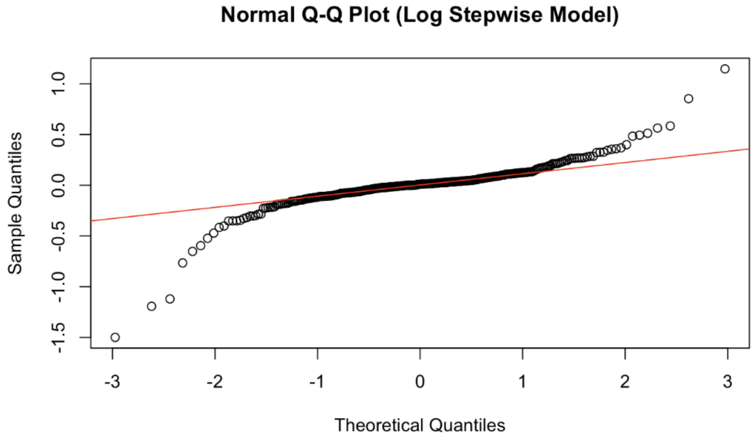
Overall, the diagnostic analysis indicates that the key regression assumptions are reasonably satisfied. While mild heteroskedasticity is present, it is addressed through the use of robust standard errors. Linearity, approximate normality, and independence appear adequate, supporting the validity of the stepwise regression results and subsequent interpretation.

Log-Transformed Model

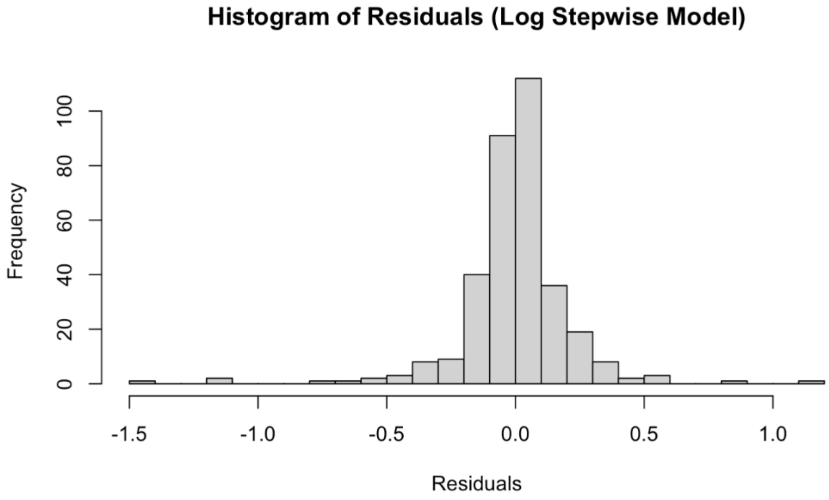
The log-transformed model is estimated to assess whether transforming the response variable improves compliance with linear regression assumptions, particularly constant variance and normality of residuals, given the right-skewed distribution of enrollment totals observed in the original model.



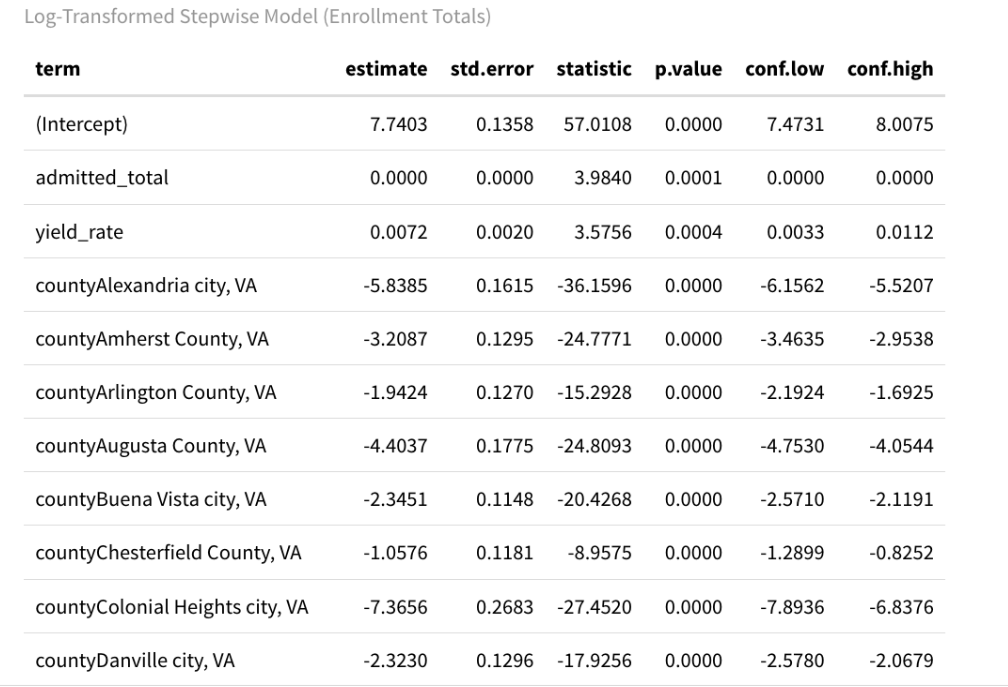
Compared to the original stepwise model, the residuals versus fitted plot for the log-transformed model shows a more uniform spread of residuals across fitted values. The funnel-shaped pattern observed in the untransformed model is substantially reduced, indicating improved homoskedasticity. Residuals remain centered around zero, supporting the linearity assumption under the transformed specification.



The Q–Q plot demonstrates improved alignment with the theoretical normal distribution relative to the untransformed model. While mild deviations remain in the extreme tails, the central portion of the distribution closely follows the reference line, indicating stronger approximate normality of residuals after transformation.



The histogram of residuals further supports this conclusion, showing a more symmetric and concentrated distribution around zero. This improvement is consistent with expectations when applying a log transformation to right-skewed count data.



Log-Transformed Stepwise Model (Enrollment Totals)

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	7.7403	0.1358	57.0108	0.0000	7.4731	8.0075
admitted_total	0.0000	0.0000	3.9840	0.0001	0.0000	0.0000
yield_rate	0.0072	0.0020	3.5756	0.0004	0.0033	0.0112
countyAlexandria city, VA	-5.8385	0.1615	-36.1596	0.0000	-6.1562	-5.5207
countyAmherst County, VA	-3.2087	0.1295	-24.7771	0.0000	-3.4635	-2.9538
countyArlington County, VA	-1.9424	0.1270	-15.2928	0.0000	-2.1924	-1.6925
countyAugusta County, VA	-4.4037	0.1775	-24.8093	0.0000	-4.7530	-4.0544
countyBuena Vista city, VA	-2.3451	0.1148	-20.4268	0.0000	-2.5710	-2.1191
countyChesterfield County, VA	-1.0576	0.1181	-8.9575	0.0000	-1.2899	-0.8252
countyColonial Heights city, VA	-7.3656	0.2683	-27.4520	0.0000	-7.8936	-6.8376
countyDanville city, VA	-2.3230	0.1296	-17.9256	0.0000	-2.5780	-2.0679

The log-transformed model retains the same core predictors as the original stepwise model, including admitted totals, yield rate, and county fixed effects. Coefficient signs and statistical significance are largely consistent with the untransformed specification, indicating that substantive relationships between enrollment and admissions pipeline variables are robust to transformation.

Under the log specification, coefficients are interpreted as approximate percentage changes in enrollment rather than absolute changes. While this interpretation can be useful, it is less intuitive for policy-facing discussions focused on enrollment counts.

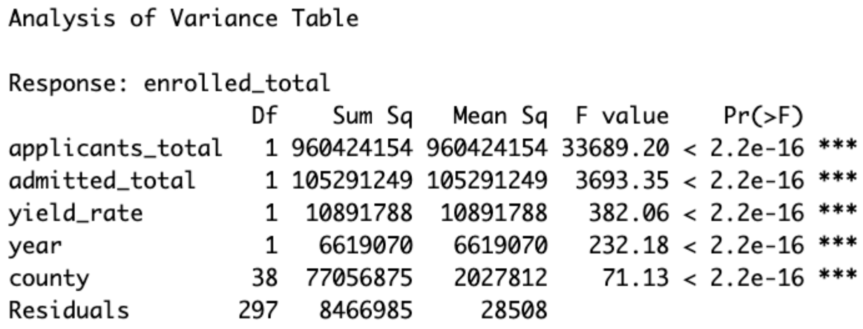
Overall, diagnostic performance improves under the log-transformed model, particularly with respect to variance stabilization and residual normality. However, the original stepwise model is retained for primary interpretation for two reasons. First, enrollment totals are naturally measured and interpreted in absolute units, making coefficient interpretation more straightforward. Second, heteroskedasticity in the original model was adequately addressed using

HC1 robust standard errors, preserving valid inference without sacrificing interpretability.

The log-transformed model therefore serves as a robustness check, confirming that the main findings are not driven by functional form, while the untransformed stepwise model remains the preferred specification for substantive conclusions.

Statistical Inference & Significance

Table: ANOVA Table for Final Stepwise Model



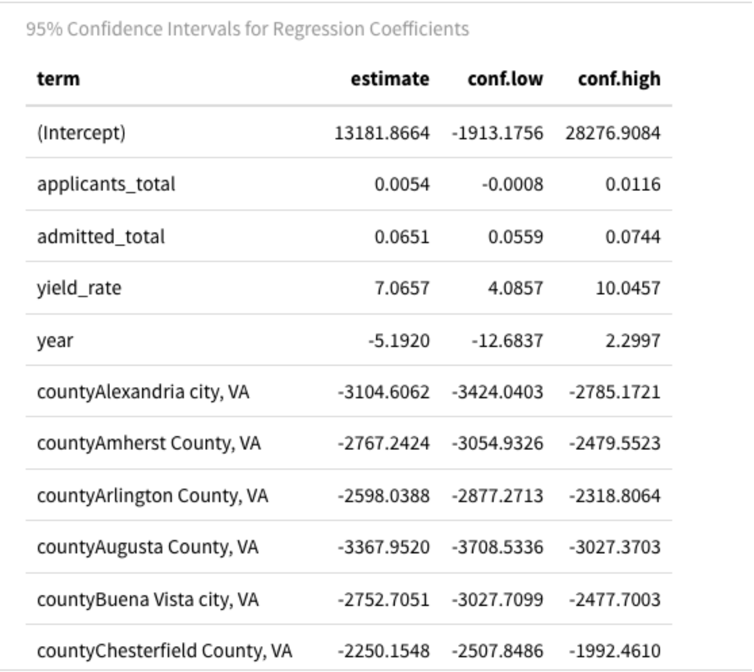
```
Analysis of Variance Table

Response: enrolled_total
                  Df    Sum Sq   Mean Sq  F value    Pr(>F)    
applicants_total   1 960424154 960424154 33689.20 < 2.2e-16 ***
admitted_total     1 105291249 105291249  3693.35 < 2.2e-16 ***
yield_rate         1  10891788  10891788   382.06 < 2.2e-16 ***
year               1   6619070   6619070   232.18 < 2.2e-16 ***
county            38  77056875   2027812    71.13 < 2.2e-16 ***
Residuals        297   8466985     28508                       
```

The overall significance of the regression model is assessed using an F-test. The null hypothesis of the F-test states that all slope coefficients in the model (excluding the intercept) are zero, implying that the model explains no variation in enrollment totals beyond random noise.

The ANOVA results strongly reject this null hypothesis. The F-statistics for key predictors (including applicants total, admitted total, yield rate, year, and county fixed effects) are extremely large, with corresponding p-values all less than 2.2e-16. This provides overwhelming evidence that the model explains a substantial portion of the variation in enrollment totals across county-year observations.

In particular, the highly significant county term confirms the importance of accounting for geographic heterogeneity, while the significance of pipeline variables supports the relevance of admissions dynamics in determining enrollment outcomes.

Table: 95% Confidence Intervals for Regression Coefficients



95% Confidence Intervals for Regression Coefficients

term	estimate	conf.low	conf.high
(Intercept)	13181.8664	-1913.1756	28276.9084
applicants_total	0.0054	-0.0008	0.0116
admitted_total	0.0651	0.0559	0.0744
yield_rate	7.0657	4.0857	10.0457
year	-5.1920	-12.6837	2.2997
countyAlexandria city, VA	-3104.6062	-3424.0403	-2785.1721
countyAmherst County, VA	-2767.2424	-3054.9326	-2479.5523
countyArlington County, VA	-2598.0388	-2877.2713	-2318.8064
countyAugusta County, VA	-3367.9520	-3708.5336	-3027.3703
countyBuena Vista city, VA	-2752.7051	-3027.7099	-2477.7003
countyChesterfield County, VA	-2250.1548	-2507.8486	-1992.4610

Confidence intervals provide information about both statistical significance and the magnitude of effects. Predictors whose 95% confidence intervals do not include zero are statistically significant at the 5% level.

Among the continuous predictors, admitted_total and yield_rate have confidence intervals that exclude zero, indicating strong and precise positive associations with enrollment totals. Specifically, increases in admitted students and higher yield rates are both associated with higher enrollment, holding other variables constant. These variables also exhibit relatively narrow confidence intervals, reinforcing their substantive importance.

In contrast, applicants_total and year have confidence intervals that include zero, suggesting that once admitted totals, yield rates, and county fixed effects are controlled for, these variables do not have statistically distinguishable effects on enrollment at conventional significance levels.

County fixed effects consistently exhibit confidence intervals far from zero, reflecting persistent and statistically significant differences in baseline enrollment levels across counties. These effects capture unobserved geographic factors such as population size, institutional presence, and regional education infrastructure.

Taken together, the inference results indicate that enrollment totals are most strongly driven by how many students are admitted and what proportion of those admitted ultimately enroll. Differences across counties also play a major role, while the total number of applicants and overall time trends contribute less once the full admissions pipeline is accounted for.

The combination of a highly significant F-test and narrow confidence intervals for key predictors provides strong evidence that the final stepwise model captures meaningful and interpretable relationships in the data.

Results Interpretation

This section interprets the estimated coefficients from the final stepwise regression model, focusing on the magnitude, direction, and substantive meaning of each key predictor.

The coefficient on applicants_total is positive but small in magnitude and only marginally statistically significant. This suggests that, holding admissions decisions, yield behavior, year, and county fixed effects constant, increases in the number of applicants are associated with only modest changes in enrollment. In practical terms, growth in applicant volume alone does not substantially translate into higher enrollment unless it is accompanied by corresponding increases in admissions or yield. This highlights the limited direct role of applicant volume once later stages of the admissions pipeline are accounted for.

Admitted_total emerges as a strong and statistically significant positive predictor of enrollment. The estimated coefficient indicates that increases in the number of admitted students are directly associated with higher enrollment totals, holding all other variables constant. This result is intuitive and reinforces the central role of admissions decisions in shaping final enrollment outcomes. Compared to applicant totals, admissions represent a more immediate and controllable lever for institutions seeking to influence enrollment size.

The yield_rate variable exhibits the largest marginal impact among the continuous predictors. A one-unit increase in yield rate is associated with a substantial increase in enrollment, even after controlling for admitted totals and county fixed effects. This finding underscores that enrollment outcomes are not determined solely by how many students are admitted, but also by how successful institutions are at converting admitted students into enrolled students. Yield rate therefore represents a particularly powerful driver of enrollment variation across county-years.

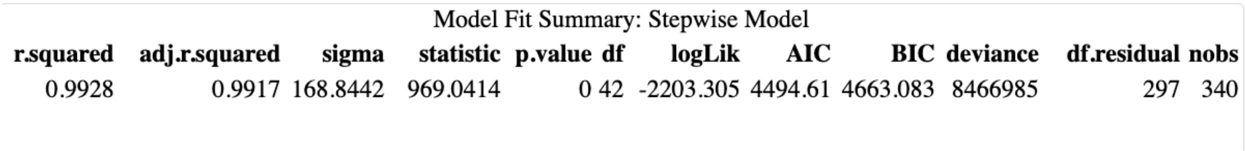
The coefficient on year is negative but not statistically significant, indicating no strong evidence of a consistent upward or downward time trend in enrollment once admissions pipeline variables and county fixed effects are included. This suggests that changes in enrollment over time are largely explained by shifts in applicant pools, admissions decisions, and yield behavior rather than by a standalone temporal trend.

The county fixed effects reveal substantial and statistically significant differences in baseline enrollment levels across counties. These coefficients capture persistent geographic disparities that remain even after controlling for admissions pipeline variables and year. Such differences likely reflect underlying factors such as population size, local education infrastructure, institutional presence, and demographic composition. The magnitude and consistency of these effects highlight the importance of accounting for geographic heterogeneity when modeling enrollment outcomes.

County coefficients are interpreted relative to the reference county omitted from the model. A negative county coefficient indicates that, all else equal, that county has lower baseline enrollment than the reference county, while a positive coefficient would indicate higher baseline enrollment. These effects reflect structural differences across counties rather than short-term fluctuations, reinforcing the appropriateness of treating county as a fixed effect in the model.

Model Performance & Fit

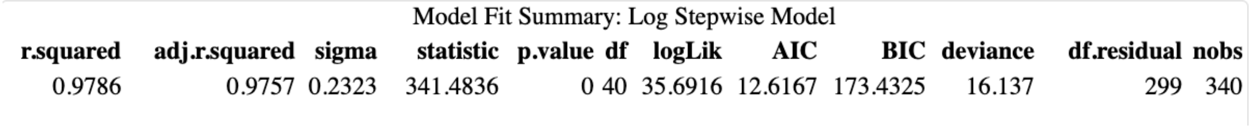
This section evaluates the overall fit and performance of the final stepwise regression model and compares it to the log-transformed alternative.



Model Fit Summary: Stepwise Model

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.9928	0.9917	168.8442	969.0414	0	42	-2203.305	4494.61	4663.083	8466985	297	340

The stepwise model explains a very large proportion of the variation in enrollment totals, with an R² of 0.9928 and an adjusted R² of 0.9917. These values indicate that the included predictors and county fixed effects collectively capture nearly all observed variation in enrollment across county-year observations. The residual standard error (σ = 168.84) reflects the average magnitude of unexplained variation in enrollment counts.



Model Fit Summary: Log Stepwise Model

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.9786	0.9757	0.2323	341.4836	0	40	35.6916	12.6167	173.4325	16.137	299	340

The log-transformed model also demonstrates strong fit, with an R² of 0.9786 and an adjusted R² of 0.9757. While these values remain high, they are slightly lower than those of the untransformed model. However, the residual standard error is substantially smaller on the log scale (σ = 0.2323), reflecting reduced heteroskedasticity and improved residual behavior.

Overall, while the log-transformed model improves adherence to regression assumptions, the original stepwise model provides superior explanatory power in terms of variance explained. As a result, the untransformed stepwise model is retained as the primary model for interpretation, with the log-transformed model used as a robustness check.

Discussion & Connection to Literature

The results of this analysis align closely with prior research on disparities across the postsecondary admissions pipeline while extending existing work by modeling multiple pipeline stages jointly.

Consistent with Reardon (2011), disparities reflected in applicant volumes appear to play a limited direct role in explaining enrollment once later pipeline stages are accounted for. While applicant totals are positively associated with enrollment, their marginal effect is relatively small compared to later stages. This supports the idea that underlying socioeconomic inequalities shape who applies to college, but application volume alone doesn’t find final enrollment outcomes.

The strong and statistically significant effect of admitted_total mirrors findings from Posselt et al. (2012), who emphasize the role of institutional filtering in shaping access to higher education. Admissions decisions emerge as a critical structural gatekeeping mechanism, with substantial influence on enrollment outcomes even after controlling for geographic differences. This suggests that institutional practices play a central role in translating applicant pools into enrolled students.

The largest marginal effect is observed for yield_rate, consistent with Hoxby and Avery

(2013), who document disparities in post-admission enrollment behavior. High yield rates reflect successful conversion of admitted students into enrolled students and capture behavioral and informational barriers that persist after admission. The prominence of yield rate in the model highlights the importance of post-admission decision-making in shaping enrollment disparities.

Overall, the post-admission stage shows the strongest association with enrollment outcomes. By modeling applicant volume, admissions decisions, and yield behavior simultaneously, this analysis extends prior work that typically examines these stages in isolation. The results demonstrate that enrollment disparities are not driven by a single stage of the pipeline, but rather emerge from compounding effects across multiple transitions, with post-admission behavior playing a particularly influential role.

Limitations & Ethical Considerations

Limitations

This analysis relies on county-level aggregation, which may obscure meaningful variation within counties and mask individual-level disparities in access to postsecondary education. As a result, findings should be interpreted as reflecting broad geographic patterns rather than individual student outcomes.

The dataset does not include individual demographic characteristics such as race, income, or first-generation status. Consequently, the analysis cannot directly assess how specific demographic factors contribute to enrollment disparities, nor can it disentangle within-county heterogeneity.

Additionally, some relationships between admissions pipeline stages are mechanical by construction. For example, enrollment is necessarily bounded by admissions, which may inflate correlations among pipeline variables. While multivariable modeling mitigates this concern, it does not eliminate structural dependence across stages.

Finally, the data are observational, limiting the ability to draw causal conclusions. The estimated associations describe relationships between pipeline stages and enrollment outcomes but should not be interpreted as causal effects.

Ethical Considerations

Care was taken to avoid deficit framing of counties with lower enrollment outcomes. Observed differences are interpreted as reflecting structural, institutional, and geographic factors rather than shortcomings of specific communities.

Interpretation emphasizes a structural perspective, situating enrollment disparities within broader systems of educational access and institutional decision-making. This framing aligns with ethical best practices by avoiding stigmatization and acknowledging the complex social contexts underlying observed patterns.

Implications & Future Research

The results suggest that yield-focused interventions may be particularly effective in increasing enrollment. Because yield rate exhibits the largest marginal association with enrollment, policies that reduce informational barriers, improve financial aid clarity, or strengthen post-admission outreach could meaningfully improve enrollment outcomes without expanding applicant pools.

The strong role of admissions totals highlights the importance of greater transparency in admissions practices. Clearer communication around admissions criteria and decision-making processes may help mitigate institutional filtering effects that contribute to persistent enrollment disparities.

Finally, the presence of large county fixed effects underscores the need for regional equity monitoring. Policymakers and educational institutions can use geographically disaggregated data to identify areas where structural barriers to enrollment remain most pronounced and allocate resources accordingly.

Future work would benefit from individual-level data, allowing for more precise analysis of how student characteristics interact with admissions pipeline stages. Incorporating financial aid variables (such as grant availability, net price, and aid offers) would further clarify the mechanisms driving post-admission enrollment decisions. Additionally, multilevel or mixed-effects models could better account for hierarchical data structures and distinguish within-county from between-county variation.

Reflection

This project strengthened my understanding of multivariable regression diagnostics, including model specification, multicollinearity assessment, and residual analysis. Interpreting assumption violations and evaluating when transformations are appropriate deepened my appreciation for the limitations of linear models.

The iterative process of model checking emphasized the importance of validation and robustness, particularly when working with real-world observational data. Overall, this project contributed significantly to my growth as a data scientist by reinforcing the value of careful modeling, transparent interpretation, and ethical analytical practice.

References

Duncan, G. J., & Murnane, R. J. (2013, December 31). *Restoring opportunity: The crisis of inequality and the challenge for American Education.* Harvard Education Press. https://eric.ed.gov/?id=ED568809

Hoxby, C. M., & Avery, C. (2012, December 6). *The missing "one-offs": The hidden supply of high-achieving, low income students*. NBER. https://www.nber.org/papers/w18586

The widening academic achievement gap between the rich and the poor: (n.d.). https://cepa.stanford.edu/sites/default/files/reardon%20whither%20opportunity%20-%20chapter%205.pdf